Into the Wide Blue Yonder with BlueGene/L

The unusual slant to BlueGene/L's cabinets is a necessary design element to keep cooled air flowing properly around each cabinet's 2,000-plus processors.

NEW vistas in numerical simulations are opening to scientists at Livermore and elsewhere with the arrival of the BlueGene/L supercomputer, a new class of highly scalable platform. IBM Research designed BlueGene/L for the Advanced Simulation and Computing (ASC) Program, a part of the Department of Energy's National Nuclear Security Administration. Last November, BlueGene/L's first segment—one-fourth the size of the ultimate machine—captured the number-one spot on the Top500 list of the world's fastest supercomputers (www.top500.org/lists/2004/11/), clocking in at 70.72 trillion operations per second (teraops) on the industry standard LINPACK benchmark. It beat the previous record holder, Japan's Earth Simulator, by a factor of two.

The second delivery of racks for BlueGene/L (originally called BlueGene/Light) arrived at Livermore's new Terascale Simulation Facility in March 2005, and the final set is scheduled for delivery this summer. The 360-teraops machine will handle many challenging ASC-related scientific simulations, including ab initio molecular dynamics; three-dimensional (3D) dislocation dynamics; and turbulence, shock, and instability phenomena in hydrodynamics. BlueGene/L is also a computational science research machine for evaluating advanced computer architectures.

BlueGene/L is a world apart from other scalable computers not only in terms of performance but also in size, appearance, and design. The machine is scaled up with a few unique components and IBM's system-on-a-chip technology developed for the embedded microprocessor marketplace. Another unusual feature is that for applications the computer's nodes are interconnected in three different ways instead of the usual one. Two principles drove the design of the hardware and software of this highly scalable machine: "keep it simple" and "divide and conquer."

Strength in Numbers

"The major difference between BlueGene/L and other computers is its scalability, that is, the sheer number of nodes we are strapping together in a single unit," says computer scientist Mark Seager, head of Advanced Technology in the Integrated Computing and Communications Department of the Computation Directorate. The most basic building block of BlueGene/L's design is the node. BlueGene/L has 65,536 nodes, compared with ASC White's 512, ASC Q's 2,048, and 1,536 in Purple, the newest mainline ASC system being constructed by IBM in Poughkeepsie, New York. "With this many nodes, achieving a reasonable hardware stability level required design simplicity, including a minimum number of chips per node," says Seager. Whereas a desktop computer can have 50 to 60 chips, a BlueGene/L node has just 10—9 memory chips and 1 compute application specific integrated circuit (ASIC) chip.

The ASIC chip is a complete system-on-a-chip that includes two IBM PowerPC 440 processors, five interconnects, and 8 megabytes of embedded dynamic random access memory. A memory controller for the nine external memory chips provides double-bit error detection and single-bit error correction. The result is a compact, low-power building block.

"The compute node has its good and bad points," notes Seager. "On the down side, it uses weak processors. On the plus side, it provides many operations per watt because it's not pushing the performance envelope. Power usage is an important consideration when you're running 65,536 nodes." The machine scales up in an orderly fashion, resulting in an extremely high-compute-density system with attractive cost performance and relatively modest power and cooling requirements.

Networking for Efficiency

Another difference between BlueGene/L and other platforms is that it has not one but three interconnects for applications: a 3D torus network, a binary-tree (combining and broadcasting) network, and a barrier network.

The 3D torus interconnect is used for high-bandwidth communication between nearest-neighbor nodes and works well on grid-based applications. This interconnect is similar to a mesh. Each node is connected to its six nearest-neighbor nodes, but the ends of the mesh loop back to the nodes, making it a torus. Seager says, "This configuration makes programming for BlueGene/L much easier than programming for a system with edges that do not have six nearest neighbors.”

Another plus is that a torus network requires far fewer cables than other types of interconnects at this scale. “When you build very large machines such as BlueGene/L, the cable issue becomes critical,” says Seager. “If we’d used a different network design, the sheer number of cables would have made building the machine impossible.”

BlueGene/L’s binary-tree network is useful for low-latency global operations that share data and synchronize programs. This interconnect determines how a highly parallel computer program “talks” to *all* the nodes quickly and efficiently. “Different ways exist to deliver a message to a large number of nodes,” says Seager. “In a binary-tree network, one node talks to two neighbors, those two talk to two of their neighbors, and so on. Getting the message out to 65,536 nodes is a very efficient process, taking only 16 tree operations, or hops.”

The binary tree can operate in broadcast mode to replicate information across the machine or in combining mode to gather data distributed across the machine into a single location. Both broadcast and combining modes are used in operations performed millions of times in real scientific applications. In BlueGene/L, the binary-tree interconnect is implemented in the hardware rather than in the software, making those hops extremely fast. Performing those operations in the hardware, says Seager, is a huge leap forward in making BlueGene/L scalable and fast.

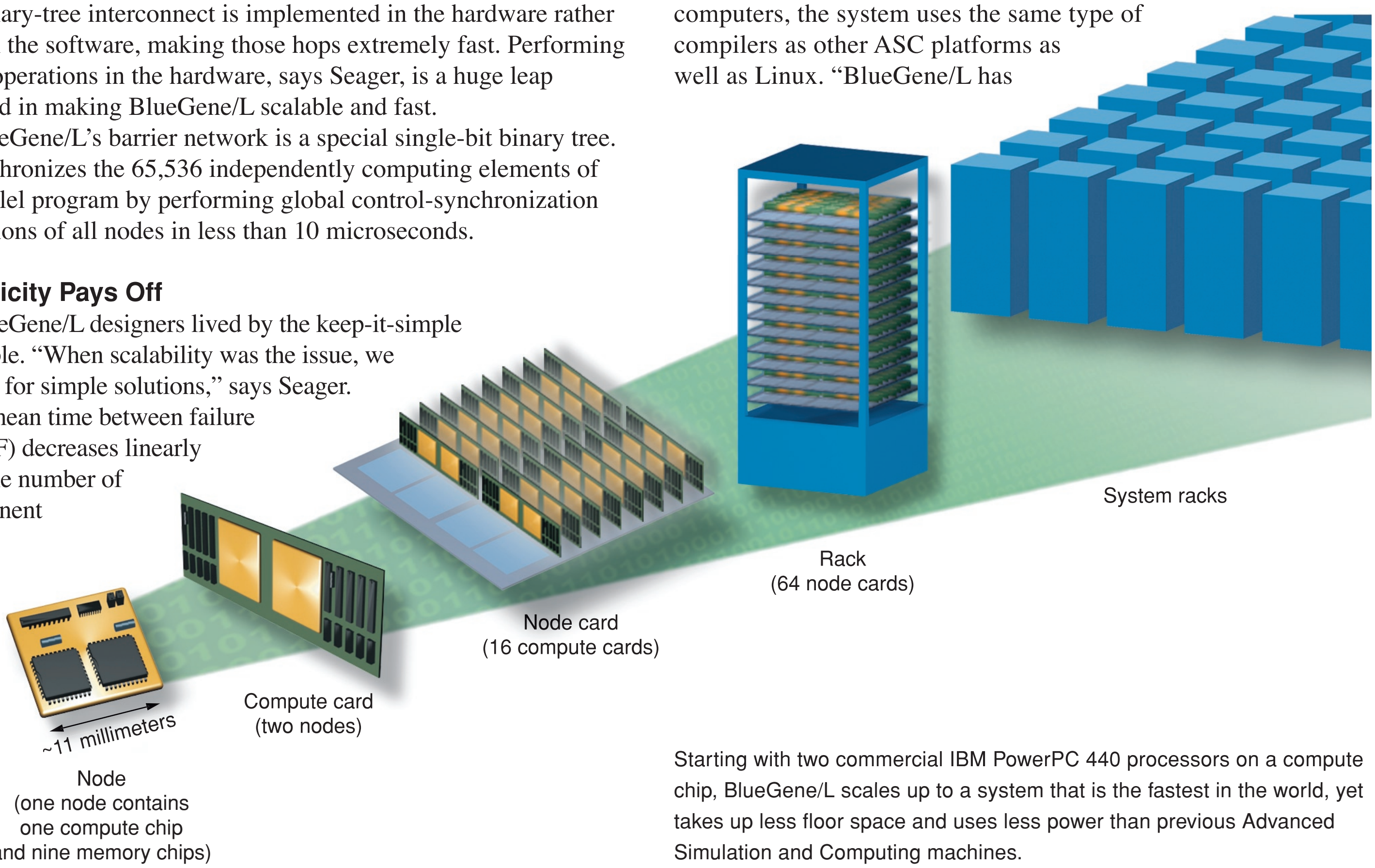
BlueGene/L’s barrier network is a special single-bit binary tree. It synchronizes the 65,536 independently computing elements of a parallel program by performing global control-synchronization operations of all nodes in less than 10 microseconds.

Simplicity Pays Off

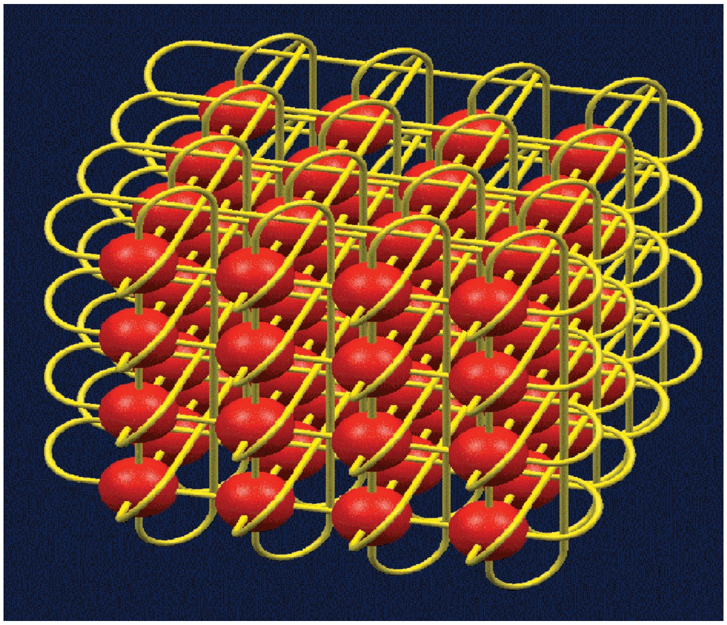
BlueGene/L designers lived by the keep-it-simple principle. “When scalability was the issue, we looked for simple solutions,” says Seager. “The mean time between failure (MTBF) decreases linearly with the number of component replications. To achieve exceptional component MTBF, we simplified component designs as much as possible.” The compute node contains the minimum number of chips, and the compute node kernel (or operating system) is also stripped down to the minimal functionality. “The compute node kernel is almost stupid, really,” says Seager.

Many functions that typically reside in an operating system, such as process migration, memory management, and file input/output (I/O), are either eliminated or pushed out to a specially designed I/O node. Sixty-four compute nodes share a single I/O node, a configuration that embodies the divide-and-conquer principle, according to Seager. The configuration divides the problem into a simple part and a complicated part. The simple part (the compute node kernel) is replicated 65,536 times, and the complicated part (the I/O node kernel) is replicated 1,024 times—all with high reliability. Pushing functions to the I/O node kept BlueGene/L’s MTBF at six-and-a-half days to the amazement of some computer scientists, says Seager. Given the number of components, they had predicted the MTBF would never rise above a couple of minutes.

The keep-it-simple philosophy also extends to the programming environment. To make BlueGene/L readily accessible to researchers programming on other ASC computers, the system uses the same type of compilers as other ASC platforms as well as Linux. “BlueGene/L has



Starting with two commercial IBM PowerPC 440 processors on a compute chip, BlueGene/L scales up to a system that is the fastest in the world, yet takes up less floor space and uses less power than previous Advanced Simulation and Computing machines.



BlueGene/L uses a three-dimensional (3D) torus network in which the nodes (red balls) are connected to their six nearest-neighbor nodes in a 3D mesh. In the torus configuration, the ends of the mesh loop back, thereby eliminating the problem of programming for a mesh with edges. Without these loops, the end nodes would not have six near neighbors.

the code development tools and an environment common to a desktop Linux system," says Seager. The new machine, as with Livermore's other Linux clusters, will also use the Lustre global parallel file system. Once BlueGene/L is integrated onto the unclassified network, users will be able to view the generated data on Livermore's Linux visualization cluster or analyze the data on other Linux clusters.

BlueGene/L and Beyond

BlueGene/L's recent record-breaking performance is just the beginning. The real acclaim will come with the scientific breakthroughs that are sure to occur as the world's most powerful computer tackles pressing science questions.

"BlueGene/L will help us to better understand the complex physics phenomena necessary to ensure the safety and reliability of the nation's nuclear deterrent," says Dona Crawford, Livermore's associate director for Computation. "This capability, in turn, is applicable to other domains, allowing us to advance multiple national agendas in science, national security, and industrial competitiveness."

Even as ASC users at Livermore and elsewhere prepare for BlueGene/L, Livermore's computer scientists are focusing on the next step. BlueGene/L, powerful as it is, still will not be powerful enough to simulate all the complexities of matter at extreme pressures and temperatures, so a petaops (1 quadrillion operations per second) computer will likely be necessary by 2008. "We have several interesting options," says Seager.

A next-generation, petaops-sized BlueGene/P beckons. Livermore is also partnering with Stanford University, which is developing a 4-petaops streaming computer as part of its work for the ASC Alliance Program. "And we're looking at using the Intel/AMD ecosystem to build a petaops machine out of mass-market commodity components," adds Seager. "The petaops era is on the horizon. When it arrives, we'll be ready."

—Ann Parker

Key Words: Advanced Simulation and Computing (ASC) Program, barrier interconnect, binary-tree interconnect, BlueGene/L, node, scalability, stockpile stewardship, supercomputer, three-dimensional torus interconnect, Top500 list.

For further information contact Mark Seager (925) 423-3141 (seager1@llnl.gov).